THE UNIVERSITY of York

CENTRE FOR HEALTH ECONOMICS

Alternative Methods to Examine Hospital Efficiency: Data Envelopment Analysis and Stochastic Frontier Analysis

Rowena Jacobs

DISCUSSION PAPER 177

ALTERNATIVE METHODS TO EXAMINE HOSPITAL EFFICIENCY: DATA ENVELOPMENT ANALYSIS AND STOCHASTIC FRONTIER ANALYSIS

Rowena Jacobs

February 2000

ABSTRACT

There has been increasing interest in the ability of different methods to rank efficient hospitals over their inefficient counterparts. The UK Department of Health has used three cost indices to benchmark NHS Trusts. This study uses the same dataset and compares the efficiency rankings from the cost indices with those obtained using Data Envelopment Analysis (DEA) and Stochastic Cost Frontier Analysis (SCF). The paper concludes that each method each has particular strengths and weaknesses and potentially measure different aspects of efficiency. Several specifications should be used to develop ranges of inefficiency to act as signalling devices rather than point estimates. There appears to be a large amount of random 'noise' in the study which suggests that there are not truly large efficiency differences between Trusts, and savings from bringing up poorer performers would in fact be very modest.

1. INTRODUCTION

Increasing emphasis is being placed on measures of efficiency in hospitals to compare their relative performance given the need to ensure the best use of scarce resources. Few studies have however assessed the consistency of efficiency rankings across different methodologies. In the UK, benchmarking has been proposed as a method whereby NHS hospitals (Trusts) could be compared to their peers and under-performing Trusts identified so that appropriate corrective actions might be taken. Several measures have been used to benchmark Trusts including NHS Efficiency Indices, NHS Performance Indicators, the Labour Productivity Index (LPI) and cost information on NHS Trusts in the form of “league tables” (Hollingsworth & Parkin, 1998). More recently, regression analysis has been applied to Trust data to develop three cost indices (the CCI, 2CCI and 3CCI) which can be used to produce productivity rankings (Söderlund & van der Merwe, 1999).

This study uses the same Department of Health cross-sectional dataset of NHS Trusts and employs two different approaches that have been more generally used to study hospital efficiency. The purpose of the study is to then compare the efficiency rankings of the three cost indices with those from other methodologies.

The first is Data Envelopment Analysis (DEA), a linear programming method which enables the measurement of efficiency consistent with the theoretically based concept of production efficiency. DEA typically examines the relationship between inputs to a production process (resources used in a hospital) and the outputs of that process (for example number of patients treated within each Healthcare Resource Group (HRG) with which that hospital deals). In this study, cost is used as the input. In other words, DEA examines the question: “By how much can cost be reduced without changing the output quantities (HRG spells, teaching, research, and so on) produced by NHS Trusts?”

The second technique for assessing efficiency that is employed is Stochastic Cost Frontier Analysis (SCF). This is an econometric technique which uses regression analysis to estimate a conventional cost function, with the difference being that efficiency of a Trust is measured using the residuals from the estimated equation. The error term is therefore divided into a stochastic error term and a systematic inefficiency term.

The basic research question then becomes: “Which of these methodologies can best be employed to measure efficiency in NHS Trusts and do the different methods produce a consistent ranking of Trusts based on these efficiency scores?”

In order to answer this research question, the study compares the efficiency rankings of NHS Trusts from the three cost indices CCI, 2CCI and 3CCI with rankings produced using DEA and SCF methodologies. The same variables and specifications are used so as to make the three methods as comparable as possible and then evaluate the efficiency rankings that each method produces. All three methods are based on a Department of Health cross-sectional dataset for NHS Trusts with a predominance of acute work, based on 1995/6 variables[1].

This paper is divided into the following sections. Section 2 examines the advantages and disadvantages of the methodologies employed in this study. Sections 3 and 4 describe the workings of the two main techniques used in this paper, DEA and SCF analysis. The following section 5 describes the various specifications used and the results obtained when comparing the different methods, while section 6 concludes with general findings on the robustness, validity and consistency of the methodologies, the application of these techniques in hospital cost analysis and their use in establishing NHS Trust efficiencies.

2. APPROACHES TO STUDY HOSPITAL EFFICIENCY

It is often argued that health care institutions are not expected to be efficient, as they do not adhere to neo-classical firm optimisation behaviour. However, given the vast amount of resources that go towards funding such institutions, there is a great and growing interest in examining efficiency in hospitals with the driving force for such concern being value for money.

Different methods to test efficiency are usually considered either parametric or non-parametric, where parametric methods assume a particular functional form (such as a Cobb-Douglas production function or a translog function) and non-parametric methods do not. An alternative taxonomy is that methods can be statistical or non-statistical, where statistical methods tend to make assumptions about the stochastic nature of the data. (Stochastic frontiers, as opposed to deterministic, allow for statistical ‘noise’.) Non-statistical methods such as DEA tend to be non-parametric (and deterministic), whereas statistical methods, based on frontier regression models tend to be parametric (and stochastic). Usually the frontier models make specific assumptions about the inefficiency term in the model which tend to be very restrictive (such as half-normal or constant inefficiency over time) (Wagstaff, 1989). SCF constructs a smooth parametric frontier which may as a result have an inappropriate technology, but accounts for stochastic error, whereas DEA constructs a piece-wise linear-segmented efficiency frontier based on best practice, with no assumption about the underlying technology but no scope for random error, making it more vulnerable to data errors. DEA has the advantage that it is able to manage complex production environments

[1] The derivation of the three cost indices (CCI, 2CCI and 3CCI), their productivity rankings and data definitions

with multiple input and output technologies like hospitals, but being a non-statistical method it does not produce the usual diagnostic tools with which to judge the goodness-of-fit of the model specifications produced.

Thus some trade-off exists between these methods. Non-statistical approaches such as DEA have the disadvantage of assuming no statistical noise, but have the advantage of being non-parametric and requiring no assumptions about the production frontier. SCF models on the other hand have the attraction of allowing for statistical noise, but have the disadvantage of being parametric and requiring strong assumptions about the inefficiency term. In fact, they have been criticised for their potential for mixing statistical noise and inefficiency (Skinner, 1994), particularly when the random error term does not obey the normality assumption.

In both methods however, some non-testable assumptions have to be made. In DEA one assumes no measurement error or random fluctuations whatsoever in output and in SCF one assumes a particular error distribution. SCF has the advantage over DEA in that it may allow for measurement error, but again inefficiency is identified from a non-testable assumption about the error distribution (Newhouse, 1994). Both methods may be vulnerable to measurement and misspecification error with dangers of omitting significant variables, the inclusion of irrelevant variables, the adoption of an inappropriate technology (in SCF), or the imposition of an inappropriate variable returns to scale assumption (in DEA) (Smith, 1997). The problem of endogeneity bias, where the inputs or resources may be endogenous, has been well-documented in regression based techniques and has generally been assumed to pose no problem for DEA. However efficiency estimates in DEA may be subject to the same bias if inefficient units using low levels of the endogenous resource are set tougher efficiency targets than equally inefficient units using more of the resource (Orme & Smith, 1996). Thus pitfalls relating to errors in the measurement of the inputs and outputs and errors in specification and estimation may largely affect both techniques.

There has been a rapid increase in the application of these methods to measure hospital efficiency. However, very few studies have examined whether applying different methods to the same data will affect sensitivity of efficiency rankings.

SCF and DEA models can be compared if certain assumptions are made, such as there are no allocative inefficiencies. SCF inefficiencies can then be compared directly to those obtained from DEA. Such a study has been done by Banker, Conrad and Strauss (1986) which paid particular attention to whether there were any similarities between the two approaches in ascertaining returns to scale and technical inefficiencies. The pattern of results on the two methods, though not identical, was generally similar. When scale and technical efficiencies

are explained more fully in Söderlund and van der Merwe (1999).

were combined for DEA, the two methods showed broadly similar efficiency scores. However, they argued that the methods might be sensitive to outliers and possible specification, measurement and data errors which could confound comparisons. Thus the verdict still seems to be out as to the degree of convergence between efficiency scores from the different techniques and their relative merits in measuring this.

3. THE DATA ENVELOPMENT ANALYSIS METHODOLOGY

Efficiency in DEA is defined as the ratio of the weighted sum of outputs of a Trust to its weighted sum of inputs (Hollingsworth & Parkin, 1998; Smith, 1998). Given *n* outputs and *m* inputs, efficiency (h_0) for hospital 0 is defined as follows:

$$\text{maximise: } h_0 = \frac{\sum_{r=1}^{p} u_r \times y_{r0}}{\sum_{i=1}^{m} v_i \times x_{i0}}$$

subject to:

$$\frac{\sum_{r=1}^{p} u_r \times y_{rj}}{\sum_{i=1}^{m} v_i \times x_{ij}} \leq 1 \qquad j = 1,\ldots, n$$

where:

y_{r0} = quantity of output r for hospital 0

u_r = weight attached to output r, $u_r > 0$, $r = 1,\ldots, p$

x_{i0} = quantity of input i for hospital 0

v_i = weight attached to input i, $v_i > 0$, $i = 1,\ldots, m$

The weights are specific to each unit so that $0 \leq h_0 \leq 1$ and a value of unity implies complete technical efficiency relative to the sample of units under scrutiny. Since the weights are not known *a priori*, they are calculated from the efficiency frontier by comparing a particular Trust with other ones producing similar outputs and using similar inputs, known as the Trust's peers. DEA computes all possible sets of weights which satisfy all constraints and chooses those which give the most favourable view of the Trust, that is the highest efficiency score.

This can be stated as a mathematical linear programming problem by constraining either the numerator or the denominator of the efficiency ratio to be equal to one. The problem then becomes one of either maximising weighted output with weighted input equal to one or minimising weighted input with weighted output equal to one (Parkin & Hollingsworth, 1997).

The input minimising programme (using duality in linear programming) which is used in this study is as follows, for hospital 0 in a sample of *n* hospitals:

minimise: $h_0 = Z$

subject to:

$$\sum_{j=1}^{n} x_{ij} \times \lambda_j \leq x_{i0} Z \qquad j = 1,\ldots, n$$

$$\sum_{j=1}^{n} \lambda_j \times y_{rj} \geq y_{r0} \qquad j = 1,\ldots, n$$

where:

$\lambda_j \geq 0, j = 1,\ldots, n$

λ_j are weights on units sought to form a composite unit to outperform j_0

The model is solved giving each Trust in the sample an efficiency score. The model computes the factor Z needed to reduce the input of hospital 0 to a frontier formed by its peers, or convex combinations of them, which produce no less output than hospital 0 and use a fraction Z of input of hospital 0. The Trust will be efficient if Z equals one. In other words a composite unit cannot be constructed which outperforms it. If Z is smaller than one, the Trust will be inefficient. The composite unit provides targets for the inefficient unit and Z represents the maximum inputs a Trust should be using to attain at least its current output (Hollingsworth & Parkin, 1998).

This paper uses the input-orientation in the DEA models described above, which essentially addresses the question “By how much can input quantities be proportionally reduced without changing the output quantities produced?” The input orientation was selected in this study, as input quantity (cost) is the primary decision variable over which Trust managers have most control. Input quantity (cost) is also examined as the source of variation in efficiency across Trusts.

The weighted combination of inputs over outputs therefore forms the production frontier. The Trusts which lie on the frontier have an efficiency score of one, using the weights of a reference Trust, and are called the ‘peers’ of the reference Trust.

DEA can be carried out with either the constant or variable returns to scale assumption (CRS or VRS). The model consistent with the CRS production frontier described above is given a further constraint in order to calculate the VRS frontier:

$$\sum_{j=1}^{n} \lambda_j = 1$$

The VRS approach produces technical efficiency scores which are greater than or equal to those obtained using CRS and is therefore probably the more flexible assumption of the underlying production technology (Coelli, 1996a).

4. THE STOCHASTIC COST FRONTIER METHODOLOGY

SCF is a statistical technique that generates a stochastic error term and an inefficiency term by using the residuals from an estimated production or cost frontier. The econometric model is typically defined to be:

$Y_i = x_i\beta + e_i$

where:

Y_i = the (logarithm of the) cost of production of the i-th firm,

x_i = a k×1 vector of (transformations of the) input prices and output of the i-th firm,

β = a vector of unknown parameters, and

e_i = the error term

The stochastic frontier approach allows the residual e to be decomposed into two parts:

$e_i = V_i + U_i$

where:

V_i = random variables assumed to be iid $N(0,\sigma_V^2)$ and independent of the U_i

U_i = non-negative random variables assumed to account for the cost of inefficiency in production, which are often assumed to be iid $N|(0,\sigma_U^2)|$ (in other words half-normal, or the absolute value of a variable distributed as $N(0,\sigma_U^2)$)

V_i is caused by stochastic noise, for example unexpected expenditures for hospital repairs or a temporarily high level of (unobservable) disease severity. The U_i is the degree of inefficiency or the distance from the cost function frontier. Although the two components of the residual can have a number of different distributions, a common assumption in the estimation procedure is that V_i is normally distributed, while U_i is often represented by a half-normal distribution. Other possible specifications include the truncated normal or exponential distribution. The cost function is then (Coelli, 1996b):

$$Y_i = x_i\beta + (V_i + U_i) \qquad i=1,...,N,$$

In this cost function the U_i now defines how far the firm operates above the cost frontier. If allocative efficiency is assumed, the U_i is closely related to the cost of technical inefficiency which may arise from managerial slack, outmoded equipment or inadequate staffing. If this assumption is not made, the interpretation of the U_i in a cost function is less clear, with both technical and allocative inefficiencies possibly involved[2].

Predictions of individual firm cost efficiencies are estimated from stochastic cost frontiers. The measure of cost efficiency relative to the cost frontier is defined as:

$EFF_i = E(Y_i|U_i, X_i)/ E(Y_i|U_i=0, X_i)$,

where Y_i is the cost of the i-th firm (Coelli, 1996b). EFF_i will take a value between one and infinity and can be defined as:

$(x_i\beta + U_i) / (x_i\beta)$

This expression for EFF_i relies upon the value of the unobservable U_i being predicted. This is achieved by deriving expressions for the conditional expectation of these functions of the U_i, conditional upon the observed value of $(V_i + U_i)$.

5. COMPARISON OF DIFFERENT METHODOLOGIES

5.1 The three cost indices (CCI, 2CCI and 3CCI)

Three separate cost indices have been developed for the Department of Health to produce efficiency rankings for Trusts in order to benchmark their performance based on their productivity scores. This analysis was based on 1995/6 data using a deterministic cost frontier regression (described more fully in Söderlund and van der Merwe, 1999). The CCI cost index is a deterministic cost index of actual divided by expected costs, where expected costs are average national costs per respective attendance and activity measures include case-mix adjusted inpatient, first outpatient and accident and emergency (A & E) attendances. 2CCI and 3CCI are long and short run indices regressed against the CCI with increasing numbers of explanatory variables. 2CCI takes factors into account such as additional adjustments for case mix, age and gender mix, transfers in and out of the hospital, inter-specialty transfers, local labour and capital prices and teaching and research costs for which Trusts might be over or under compensated. The 3CCI makes additional adjustments over and above those in the 2CCI for hospital capacity, including number of beds, and number of sites, scale of inpatient and non-inpatient activity and scope of activity. It therefore tries to capture institutional

[2] Any failure in optimisation, whether technical or allocative, will show up as higher cost. The computation is dependent on the inputs chosen and whether they are allocatively efficient. Thus, a producer may be operating technically efficient by a production function, but show up as inefficient with respect to a cost function. Therefore it has been argued that the interpretation of the one-sided error on the cost side as a measure of technical inefficiency is only appropriate if the measure is defined in terms of costs, rather than output. Thus one should measure efficiency by costs rather than outputs (Greene, 1993). Inefficiency is therefore often interpreted as “cost inefficiency”, the total of both technical and allocative inefficiency.

characteristics amenable to change in the long, but not the short run (Söderlund and van der Merwe, 1999). The variables in these benchmarking regressions are shown in Table 1.

Table 1: Variables in benchmarking regressions, 1995/6

Dependent variable	
CCI	Cost index
In short and long run estimates (2CCI and 3CCI)	
TRANSIPP	Transfers in to hospital per spell
TRANSOPP	Transfers out of hospital per spell
EMERGPP	Emergency admissions per spell
FCEINPP	Finished consultant episode inter-specialty transfers per spell
OPNPP	Non-primary outpatient attendances per inpatient spell
EMERINDX	Standardized index of unexpected emergency admissions/total emergency admissions
EP_SPELL	Episodes per spell
HRGWTNHS	HRG weight, case mix index
PROP15U	Proportion of patients under 15 years of age
PROP60P	Proportion of patients 60 years or older
PROPFEM	Proportion of female patients
STUDENPP	Student whole time teaching equivalents per inpatient spell
RESEARPC	Percentage of total revenue spent on research (estimated 1995)
MFF_COMB	Market forces factor – weighted average of staff, land, buildings and London weighting factors.
In short run model only (3CCI)	
HESSPNHS	Total inpatient spells by NHS patients
TOTOP1	Total primary outpatient attendances (NHS patients)
A_E1	Total primary A & E attendances (NHS patients)
AVBEDS	Average available beds
HEATBED	Heated volume per bed
SITES50B	Sites with more than 50 beds
ITINDX	Scope / specialization index, information theory index

The regressions were run on a full sample with all Trusts and a trimmed sample which excluded outlier data representing atypical providers. Productivity scores were then derived from the residuals of the regressions. All three productivity scores were standardised to give a national average of approximately 1 with higher scores indicating less efficient providers.

Trusts were ranked against one another on these three indices. The basic descriptive statistics for these indices are shown in Table 2.

Table 2: Efficiency rankings from benchmarking regression model

Key		Mean	Std Dev	Min	Max
CCI (Cost index)	The unadjusted cost index	1.015	0.218	0.531	2.166
2CCI (LR index)	Partially adjusted cost index - adjusted only for factors that would be exogenous in the long run	1.010	0.151	0.606	1.819
3CCI (SR index)	Fully adjusted cost index using all regressors	1.003	0.115	0.581	1.603

5.2 DEA model specifications

DEA model specifications were developed based on the same variable set as the above regression models, in order to produce efficiency scores that would be comparable to the three indices.

The DEA methodology uses the relationship between inputs and outputs to establish efficiency scores. The input used in the DEA analysis in each case, is the cost index CCI, the dependent variable from the regression model. The cost index as the input is unusual as it already represents a case-mix adjusted efficiency measure, but it was kept as the input to maintain as much consistency as possible with the original regression and to keep all variables in ratio form.

Only ratio or index variables are used and so scale of activity variables (number of spells, beds, first outpatient and first A & E attendances) from the regression model were excluded. The only non-ratio data variable left in the model was SITES50B, which effectively means that Trusts are compared to other Trusts in the analysis with the same number of sites as themselves or with others on more sites.

With cost as the input, various specifications were attempted using the selection of variables from the benchmarking model that are listed in Table 3. All outputs for the DEA were transformed so as to be positively related to efficiency, for example regression coefficients that were hypothesised to be negative had to be transformed (Transfers out per spell became [1- (Transfers out per spell)]).

Table 3: DEA model variables used from benchmarking model

Inputs	
CCI	Cost index
Outputs	
EP_SPELL	Episodes per spell
TRANSIPP	Transfers in per spell
1-TRANSOPP	1 minus Transfers out per spell
EMERGPP	Emergencies per spell
FCEINPP	Specialty finished consultant episode (FCE) transfers per spell
OPNPP	Outpatient attendances per spell
EMERINDX	Emergency index / unpredictability of emergencies
PROP60P	Proportion of patients over 60 years
PROP15U	Proportion of patients under 15 years
PROPFEM	Proportion of female patients
STUDENPP	Student whole-time equivalents (WTEs) per spell
RESEARPC	Proportion of revenue from research
MFF_COMB	Market forces factor
HEATBED	Heated volume per bed in cubic meters
ITINDX	Scope / specialisation (Information Theory) index
SITES50B	Number of sites with more than 50 beds
HRGWTNHS	HRG weight, case mix index

Five model specifications were employed using the above-listed variables. Although any number and combination of variables could have been included as outputs, the following specifications were used in order to maintain some theoretical grounding and reasoning for their inclusion. Specification 1 uses all the outputs listed above and all the other specifications are therefore nested within the first one. Specification 2 uses only the variables from the benchmarking regression to obtain the short-run efficiency index 3CCI, the scope and capacity variables. The third specification uses the variables that were highly significant in the benchmarking regression models used to produce the cost indices (both the full model and the trimmed model with the outliers excluded). Specification 4 includes those variables for which there was some *a priori* hypothesis that they were positively correlated with cost and specification 5 is the variables that were significant in the full model of the benchmarking regression (including the outliers). In each case, each Trust's performance is assessed on those outputs that are included. The 5 specifications are shown in Table 4.

Given the nature of the data (ratio / proportional), a variable returns to scale (VRS) model was run, which because of the use of ratios, effectively implies an underlying constant returns to scale (CRS) technology (Fernandez-Castro & Smith, 1994). An input-orientation was selected, as the question of concern was by how much cost (the input) could be reduced in each Trust, while still producing the same outputs.

While the original benchmarking regression derived outlier-excluded efficiency estimates to reduce the extremity of scores and for the outliers only used full sample estimates, DEA calculates the efficient frontier and efficiency scores by focusing on Trusts with outstanding performance in any dimension or set of dimensions. Extreme values or outliers which are usually more suspicious and thus purged in other methods, are heavily relied on in DEA and thus included. DEA as a method does not smooth them away and thus the specifications were based on the full sample of Trusts.

Table 4: DEA model specifications for benchmarking model

Specification 1		**Specification 2**	**Specification 3**	**Specification 4**	**Specification 5**
Inputs					
CCI		CCI	CCI	CCI	CCI
Outputs					
EP_SPELL TRANSIPP 1-TRANSOPP EMERGPP FCEINPP OPNPP EMERINDX PROP15U PROP60P PROPFEM STUDENPP RESEARPC MFF_COMB HEATBED ITINDX SITES50B HRGWTNHS		HEATBED ITINDX SITES50B	OPNPP EP_SPELL PROP60P PROPFEM RESEARPC MFF_COMB HEATBED SITES50B	TRANSIPP 1-TRANSOPP OPNPP EP_SPELL EMERINDX HRGWTNHS STUDENPP MFF_COMB HEATBED ITINDX SITES50B FCEINPP	OPNPP EP_SPELL EMERINDX PROP15U PROP60P PROPFEM RESEARPC MFF_COMB HEATBED ITINDX SITES50B
Results					
Inefficient	122	224	211	173	201
Efficient	110	8	21	59	31
30th decile	0	3	1	0	0
40th decile	0	13	2	1	1
50th decile	0	63	12	1	3
60th decile	1	96	63	8	34
70th decile	24	46	92	50	84
80th decile	39	3	38	77	68
90th decile	58	0	3	36	11
Mean	0.936	0.645	0.754	0.874	0.801
Std dev	0.081	0.111	0.116	0.104	0.110
Min	0.672	0.331	0.384	0.471	0.483
Max	1.000	1.000	1.000	1.000	1.000

As seen in Table 4, the full set of variables used in specification 1 produced the higher efficiency scores with a mean of 0.94 and a lower standard deviation than the other specifications. In general, as the other specifications were nested within number 1, the efficiency scores increased as more variables were added. A higher number of Trusts also fell on the efficiency frontier in specification 1 than in any of the other specifications, which is to be expected. As more variables are added, efficiency scores increase, variability decreases and a greater number of Trusts end up on the efficient frontier with scores of 1, thus rendering the specification less discriminating.

The different specifications also serve in some sense as a sensitivity analysis, as the scores remain relatively consistent when parameters are removed and then added again.

Because one has no diagnostic tools with which to choose the best model specification, some general rules of thumb apply. The most important criterion for selecting one of these specifications is whether the model is consistent with theory and in some way theoretically justifiable. Another useful criterion is the number of efficient units. *Ceteris paribus*, the fewer the better, although there should be enough peers available to make useful comparisons. The distribution of efficiency scores makes for another useful criterion. The wider the better, *ceteris paribus* (Hollingsworth & Parkin, 1998). For these reasons specification 5 was selected as a good model. All the outputs in this specification were highly significant variables in the full model of the original benchmarking regression and this was chosen in this case over the trimmed model as DEA is based on outlier analysis.

Figure 1 shows the frequency of the distribution of efficiency scores for the 5 DEA specifications and highlights again that specification 1 produces the higher efficiency scores while specification 5 produces a spread of efficiency scores that are more average.

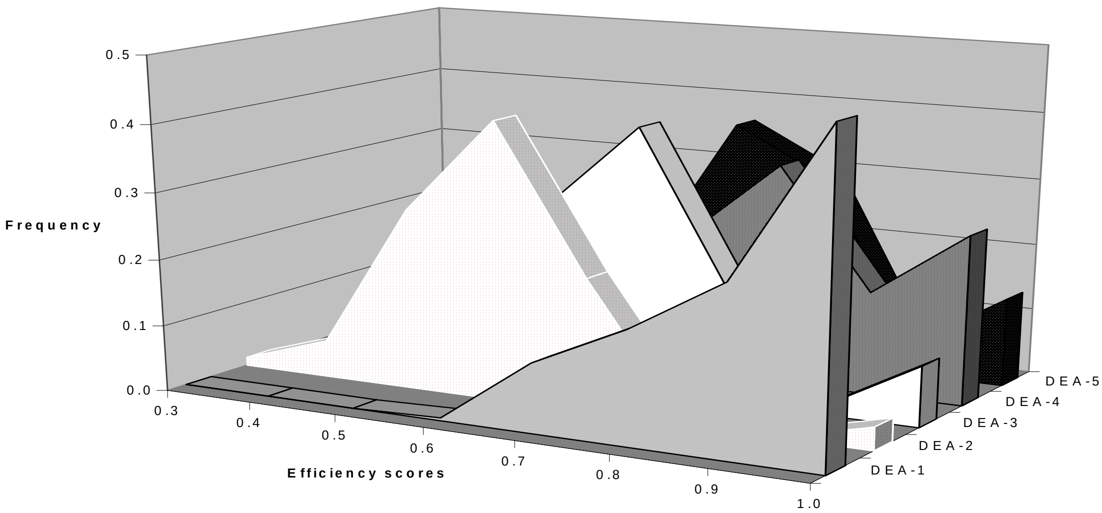

Figure 1: Distribution of efficiency scores for 5 DEA specifications

5.3 Comparison of three cost indices and DEA efficiency scores

In order to make the comparison with DEA possible, the 3 cost indices had to be re-scaled so that the most efficient Trust would rank as one and inefficient Trusts as less than one. Because the cost indices are derived from a deterministic regression, all the re-scaling does is to make 1 Trust 100 percent efficient and all others are scored relative to the most efficient one. The descriptive statistics for these (re-scaled) efficiency indices are given in Table 5.

Table 5: efficiency rankings from benchmarking model to be compared to DEA efficiency scores

Key	CCI	2CCI	3CCI
Mean	0.541	0.611	0.587
Std Dev	0.092	0.081	0.069
Min	0.245	0.333	0.362
Max	1.000	1.000	1.000
20th decile	5	0	0
30th decile	8	1	4
40th decile	46	12	11
50th decile	120	126	84
60th decile	47	84	112
70th decile	4	4	18
80th decile	1	4	2
90th decile	0	0	0
Efficient	1	1	1

Note: CCI = Efficiency rank based on cost index, 2CCI = Efficiency rank based on long run index, 3CCI = Efficiency rank based on short run index

Figure 2 shows graphically the frequency distribution of these re-scaled cost indices.

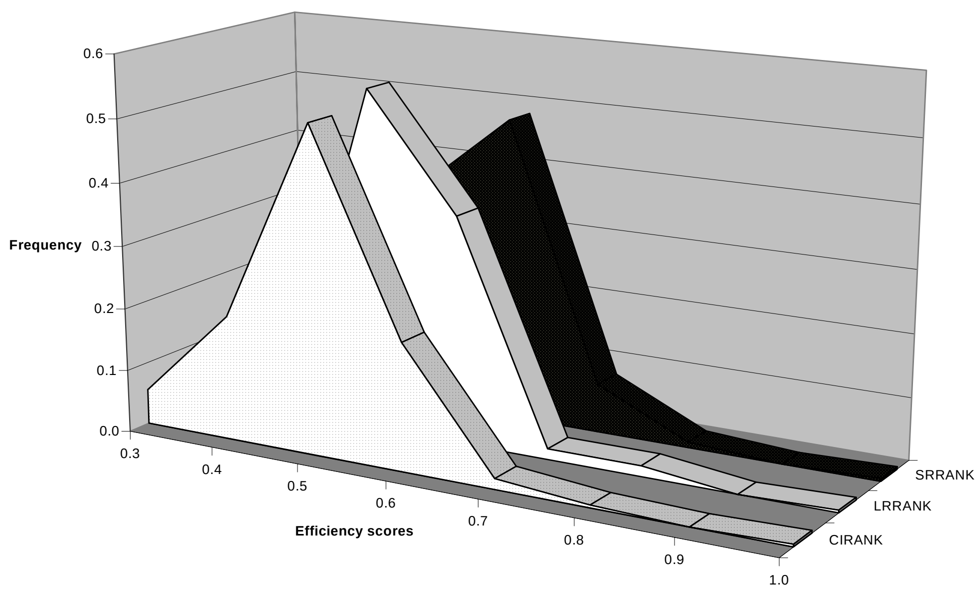

Figure 2: Distribution of efficiency scores for cost indices from benchmarking model

All three indices appear to have more skewed distributions than the 5 DEA specifications shown in Figure 1 and there does not appear to be as much variability between them as there was between the 5 DEA distributions.

Table 6 shows the correlations between the cost indices and the efficiency scores from the 5 DEA specifications. (Rank correlations were also calculated and showed very little difference to the results in Table 6 though they are not reported here.)

Table 6: Pearson correlation matrix of benchmarking regression and DEA efficiency ranking scores

	CCI	2CCI	3CCI	DEA-1	DEA-2	DEA-3	DEA-4	DEA-5
CCI	1.0000							
2CCI	0.7609	1.0000						
3CCI	0.6076	0.7967	1.0000					
DEA-1	0.0420	0.2052	0.2862	1.0000				
DEA-2	0.5358	0.3345	0.4213	0.2298	1.0000			
DEA-3	0.3856	0.4739	0.4971	0.3729	0.6340	1.0000		
DEA-4	0.0887	0.3015	0.3947	0.7575	0.3513	0.5372	1.0000	
DEA-5	0.3397	0.4467	0.4741	0.4722	0.6062	0.8352	0.6149	1.0000

Note: DEA-1 = DEA model specification 1, DEA-2 = DEA model specification 2, DEA-3 = DEA model specification 3, DEA-4 = DEA model specification 4, DEA-5 = DEA model specification 5

The correlations in Table 6 suggest a high degree of correlation between the three cost indices. They also suggest a high degree of correlation between most of the 5 DEA specifications. However, the relationship between the three cost indices and the DEA efficiency scores is generally lower, although they are all the right direction. In specifications 2, 3 and 5 reasonable correlations are achieved of around 0.5. However, there appear to be some major anomalies for individual Trusts. The techniques do not appear to be measuring efficiency related to cost in entirely the same way and the relationship does appear to be specification sensitive.

As mentioned, the DEA model relies heavily on outliers and does not smooth them away. In the benchmarking regression the outliers were excluded when calculating the efficiency scores for the non-outlier Trusts. This may partly explain some of the differences in the scores.

Correlations are not however an entirely satisfactory way to examine the changes in efficiency scores across different specifications, as they do not show what happens to individual Trusts' scores. Table 7 examines the relationship between the efficiency scores in specifications 2, 3 and 5 of the DEA method and those of the cost index CCI when grouped into deciles. Table 7 therefore shows how far out the scores actually lie from one another in the three DEA specifications compared to those of the CCI. If one had a perfect correlation, all scores would lie across the diagonal, the further away from the diagonal, the less agreement there is between the efficiency measures.

Table 7: Percentage of Trust efficiency scores that fall into each decile with respect to CCI for DEA-2, DEA-3 and DEA-5 respectively

CCI										
		20%	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
	20%									
	30%		**1%** **0%** **0%**							
	40%		1% 0% 0%	**4%** **0%** **0%**						
	50%		1% 0% 1%	11% 4% 0%	**14%** **0%** **0%**					
DEA-2 **DEA-3** **DEA-5**	**60%**		0% 1% 1%	3% 9% 9%	26% 16% 5%	**12%** **0%** **0%**				
	70%			1% 3% 6%	10% 26% 29%	7% 11% 1%	**1%** **0%** **0%**			
	80%			0% 2% 2%	0% 5% 13%	0% 8% 14%	0% 1% 0%			
	90%				0% 1% 1%	0% 0% 3%	0% 0% 1%			
	100%	0% 1% 2%	0% 0% 1%	0% 2% 3%	1% 3% 3%	1% 2% 3%				**0%** **0%** **0%**

Note: CCI = Efficiency rank based on cost index, DEA-2 = DEA model specification 2, DEA-3 = DEA model specification 3, DEA-5 = DEA model specification 5

Table 7 shows that the efficiency scores are always higher for DEA than for the cost index (CCI) and as such the scores never cross the diagonal. For DEA specification 2, which had a correlation of 0.54 with the cost index, the scores lie relatively close to the diagonal, suggesting that a large proportion (77 percent) of the scores do not lie more than one decile away from each other. For DEA specification 5 there are large discrepancies in the efficiency scores compared to the cost index. There is a huge difference in Trust efficiencies across the two measures and 92 percent of the scores are more than one decile apart. For DEA specification 3 the corresponding figure is 63 percent of Trusts whose efficiency score changes by more than one decile.

However, if one argues that the DEA scores are measuring cost efficiency in the same way as the cost index, but are just generally higher, one could hypothetically shift all scores up by one decile in Table 7 (thus effectively subtracting 10 percent off all DEA scores) to achieve a greater degree of agreement between the measures. This would mean that for DEA specification 2, only 3 percent of scores would be more than one decile out, for DEA specification 3, 20 percent of scores would differ by more than one decile with the cost index, while for DEA specification 5, 38 percent of scores would still shift by more than one decile. Thus if the scale of measurement were more similar across the two methods, there would be a large improvement in the correlation between the measures. Correlation as a method may therefore not be adequate to capture the shifts in efficiency across different methods and scaling may be part of the problem.

5.4 Comparison of SCF efficiency scores

The regression-based technique SCF analysis was used to validate and compare the efficiency rankings of the 5 DEA specifications. The five specifications were exactly replicated with the cost index (input) once again as the dependent variable and the outputs specified in each, as the regressors. A half-normal distribution was assumed for the error term. This technique produces cost efficiency scores which then need to be inverted so as to be comparable to the DEA scores. Table 8 gives the basic descriptive statistics for the efficiency scores of the 5 specifications.

Table 8: Descriptive statistics for stochastic cost frontier model efficiency scores

	Specification 1	**Specification 2**	**Specification 3**	**Specification 4**	**Specification 5**
Mean	0.865	0.831	0.842	0.876	0.867
Std dev	0.067	0.079	0.082	0.052	0.064
Min	0.600	0.523	0.507	0.652	0.617
Max	0.977	0.977	0.978	0.976	0.977
50th decile	0	7	3	0	0
60th decile	4	8	8	2	5
70th decile	34	41	56	15	34
80th decile	115	147	100	129	115
90th decile	79	29	65	86	78

Note: Specifications identical to DEA model specifications

Figure 3 displays the distribution of the efficiency scores found with stochastic frontier analysis after being converted so as to be comparable to the 5 DEA specifications.

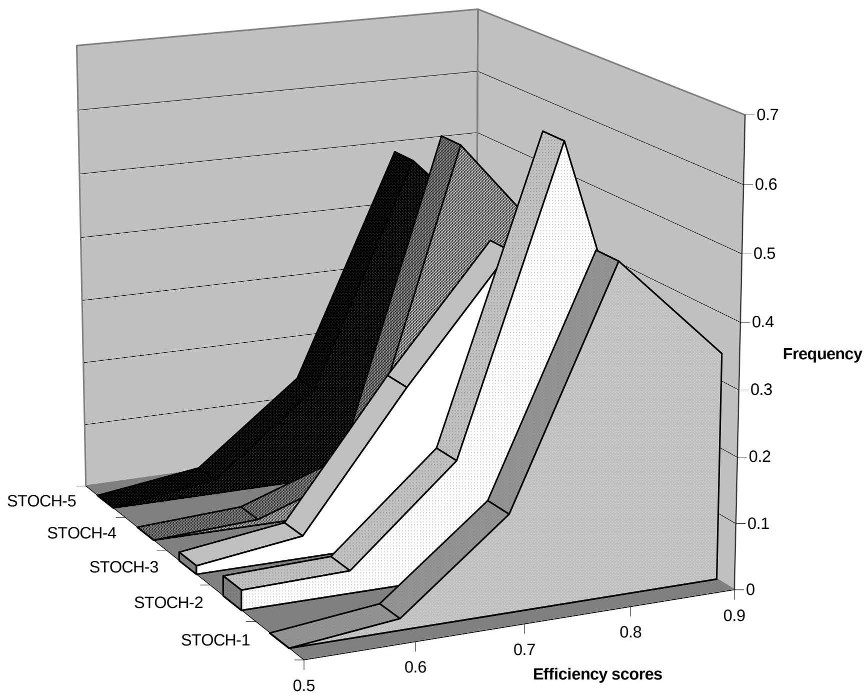

Figure 3: Distribution of efficiency scores for 5 stochastic frontier model specifications

Figure 3 shows a much narrower distribution of scores in general compared to DEA with relatively little variability across the 5 specifications.

Table 9 shows the correlations between the 5 DEA efficiency ranking scores, the 5 SCF efficiency ranking scores and also the three original benchmarking regression efficiency indices.

There is a high degree of correlation between the two regression based techniques (the benchmarking regression and the stochastic frontier analysis). Correlations are around 0.7. Correlations within the 5 stochastic specifications are also high, around 0.8. As seen before, correlations within the 3 regression indices is high, around 0.7. While they are also quite high within the 5 DEA specifications, it is generally more variable. However, the correlations between the DEA specifications and the other two techniques are generally lower and more disappointing. It is worth noting though the correlations between the two methods within certain specifications (along the diagonal) which are relatively high, especially for specifications 2, 3 and 5 which are between 0.59 and 0.63. Within the same specifications there may therefore be more agreement, but across different specifications and methods, the correlations fall.

Table 9: Pearson correlation matrix of benchmarking regression, DEA and stochastic cost frontier efficiency ranking scores

	CCI	2CCI	3CCI	DEA-1	DEA-2	DEA-3	DEA-4	DEA-5	Stoch-1	Stoch-2	Stoch-3	Stoch-4	Stoch-5
CCI	1.0000												
2CCI	0.6076	1.0000											
3CCI	0.7609	0.7967	1.0000										
DEA-1	0.0420	0.2052	0.2862	1.0000									
DEA-2	0.5358	0.3345	0.4213	0.2298	1.0000								
DEA-3	0.3856	0.4739	0.4971	0.3729	0.6340	1.0000							
DEA-4	0.0887	0.3015	0.3947	0.7575	0.3513	0.5372	1.0000						
DEA-5	0.3397	0.4467	0.4741	0.4722	0.6062	0.8352	0.6149	1.0000					
Stoch-1	0.5614	0.7563	0.7196	**0.4274**	0.4667	0.5946	0.5166	0.5756	1.0000				
Stoch-2	0.8680	0.7293	0.6326	0.0957	**0.6209**	0.4231	0.1831	0.4038	0.6354	1.0000			
Stoch-3	0.7064	0.8507	0.6886	0.2154	0.4318	**0.5975**	0.3165	0.4852	0.8297	0.6917	1.0000		
Stoch-4	0.5827	0.7355	0.7310	0.4192	0.4835	0.6583	**0.5543**	0.5998	0.8763	0.6815	0.8065	1.0000	
Stoch-5	0.5806	0.7535	0.6878	0.3399	0.5195	0.6557	0.4633	**0.6343**	0.9496	0.6535	0.8731	0.8217	1.0000

Note: DEA-1 = DEA model specification 1, DEA-2 = DEA model specification 2, DEA-3 = DEA model specification 3, DEA-4 = DEA model specification 4, DEA-5 = DEA model specification 5
Stoch-1 = Stochastic cost frontier model specification 1, Stoch-2 = Stochastic cost frontier model specification 2, Stoch-3 = Stochastic cost frontier model specification 3, Stoch-4 = Stochastic cost frontier model specification 4, Stoch-5 = Stochastic cost frontier model specification 5

As stated earlier therefore, the techniques do not appear to be measuring efficiency related to cost in entirely the same way and the relationship does appear to be sensitive to how the models are specified.

Table 10 shows the movement in efficiency scores between the different methods used, in particular with relation to the stochastic cost frontier specification 5. It is compared to the cost index, to another stochastic cost frontier specification, number 2 and to DEA specification 5. These had correlations with the stochastic cost frontier specification 5 of 0.58, 0.65 and 0.63 respectively. The efficiency scores are grouped in deciles with respect to the cost frontier specification 5 and all scores on the diagonal would represent a perfect correlation.

Table 10: Percentage of Trust efficiency scores that fall into each decile with respect to STOCH-5 for CCI, STOCH-2, and DEA-5 respectively

		CCI, STOCH-2, DEA-5								
		20%	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
	20%									
	30%									
	40%									
STOCH-5	**50%**									
	60%		1%	1%	0%					0%
			0%	0%	1%					0%
			0%	0%	1%					1%
	70%	1%	2%	7%	5%	0%	**0%**	0%		0%
		0%	0%	0%	1%	2%	**8%**	3%		0%
		0%	0%	0%	0%	9%	**3%**	0%		1%
	80%			9%	34%	6%	0%	**0%**	0%	0%
				0%	1%	1%	8%	**38%**	2%	0%
				0%	0%	4%	30%	**13%**	0%	3%
	90%			3%	13%	15%	2%	0%	**0%**	0%
				0%	0%	0%	2%	21%	**10%**	0%
				0%	0%	0%	3%	16%	**5%**	9%
	100%									**0%**
										0%
										0%

Note: Stoch-5 = Stochastic cost frontier model specification 5, CIRANK = Efficiency rank based on cost index, Stoch-2 = Stochastic cost frontier model specification 2, DEA-5 = DEA model specification 5,

The results in Table 10 show that mostly the cost frontier has higher scores than the others and as such most scores fall to the left and below the diagonal. In particular for the cost index CCI, listed first, no scores lie to the right of, or indeed on the diagonal. In fact all scores lie more than one decile away and are quite disparate from the stochastic frontier efficiency scores. For stochastic cost frontier specification 2, only 5 percent of the scores lie more than a decile apart, thus there is a high degree of congruence between them. There is a relatively high degree of agreement between DEA specification 5 and stochastic cost frontier specification 5 and only 12 percent of them fall more than 1 decile away from each other. Thus although the correlation of DEA-5 with STOCH-5 (0.63) would appear only marginally higher than that for STOCH-5 with the cost index (CCI) (0.58), there is clearly a higher consistency between the former results as shown by the degree to which the scores shift across deciles.

This would seem to suggest that although the different methods, for example DEA and stochastic frontier analysis, appear to be somewhat inconsistent, they do have congruity within specifications and may complement one another.

6. CONCLUSIONS AND FUTURE RESEARCH

This study has focused on the strengths and weaknesses of different research methods and the consistency and robustness of the DEA and SCF techniques in particular.

Caution is warranted against literal interpretations of Trust efficiency scores and rankings obtained as it is evident that some inconsistency exists across the different methodologies. The different efficiency scores should not be interpreted as accurate point estimates of efficiency, but might more usefully be interpreted as indicating general trends in inefficiency for certain Trusts. It should be noted that the DEA technique necessarily chooses the weights that will put the Trust in the best possible light, thus generating the best score possible. As such, the DEA results err on the conservative side.

Sensitivity analysis was carried out within the DEA and SCF models by changing the model specifications (omitting and including different variables) and testing for the robustness of the results, as done in other studies (Valdmanis, 1992). While the models proved to be robust in this respect, there was some inconsistency across the different methodologies. Reasonable correlations might have suggested convergent validity but these were at best modest across the different techniques. Reasons that have been proposed for this include the way outliers have been treated in the different methods and the fact that correlations may not necessarily be the best way to examine the relationship between sets of scores.

Another possible reason for the lack of agreement across the different methods, is that there appears to be a large amount of random ‘noise’ in the study which could potentially be mistaken for inefficiency. In other words, the actual differences in efficiency between Trusts is not that great and efficiency improvements and cost savings from bringing up poorer performing Trusts would in fact be very modest. The range of efficiency scores within each method were indeed quite narrow and differences across methods may therefore contain mostly random ‘noise’. This is a particularly important consideration in DEA which is highly dependent on outlier data.

Ultimately it would be useful for this paper to be able to say something about which methodology is deemed best and how decisions should be made about model specifications. It is fairly understandable that there would be a greater degree of congruence between the regression based techniques and correlations amongst these efficiency scores were naturally impressively high. Similarly, it is fairly understandable that correlations between DEA scores and the other two methods were fair, given the different variables included and the different ways outliers were dealt with. One might thus erroneously wish to conclude that the stochastic frontier method holds up as being more consistent. However, each of the methods does have unique strengths and weaknesses and potentially measures slightly different aspects of efficiency. By allowing different Trusts to assign different weights to outputs, DEA addresses the issue of technical efficiency. The inefficiency measured by stochastic frontier analysis may be a combination of technical and allocative inefficiency and without further assumptions the method is unable to separate the two sources (Kooreman, 1994). The distinction between allocative and technical efficiency is important, as they require different policy responses.

As with previous studies that have attempted to compare these methods (Banker, Conrad & Strauss, 1986), it is argued that they may be sensitive to outliers which certainly exist in the benchmarking model, as well as possible specification, measurement and data errors. The point estimates of inefficiency in either method are indeed sensitive to assumptions. However, ultimately when several specifications were used, general trends could be discerned as to which Trusts usually came out as being more efficient and which ones generally emerged as inefficient. It is therefore imperative that several specifications be employed to gauge an overall picture of efficiency. Given cross-sectional data, these techniques are certainly the best state of the art and most accurate available and where possible should be used in conjunction with one another as the two techniques are complementary in many respects. A grouping of relative ranges of inefficiency may be much less sensitive to technical and data issues than a point estimate of inefficiency. Both methods serve as signalling devices. The actual degree of inefficiency and the policy response will depend on the Trust’s circumstances and appropriate action should only be taken after more detailed investigation. While these methods prove useful diagnostic tools it would be inappropriate to base funding and resource decisions entirely on the back of the efficiency estimates arrived at (Hadley & Zuckerman, 1994).

Ultimately, data accuracy is paramount to any such analysis as inaccurate data in the DEA methodology will affect not only that Trust’s efficiency rating but also potentially the efficiency ratings of other Trusts as well. Certain improvements in the data would have potentially improved the models in this paper which were essentially constrained to some degree by data availability. Future research should also consider ways to improve models through the possible inclusion of some alternative variables. Omitted variables may have biased not only the DEA but also the SCF results as the coefficients of the included variables in the frontier analysis may also lead to biased estimates of inefficiency scores (Dor, 1994; Vitaliano & Toren, 1994).

While the models did take account of case-mix in so far as using HRG weights, these do not always adequately account for severity of cases. Certain proxies such as transfers of patients between hospitals and finished consultant episodes (FCEs) per spell, do give some indications of severity, but improved measures relating to severity would enhance the modelling. It should be noted however that one study found the inclusion of case-mix (as a weighting device and as a separate output) made no statistical difference to hospital performance measurement in DEA (Grosskopf & Valdmanis, 1993). This paper argues that case-mix (and severity) should be included in the analysis both for sound theoretical reasons and for credibility purposes. Given that hospital output is far from homogeneous, taking account of case-mix is the best known way to account for this heterogeneity. Many effective ways exist with which to measure severity and their application in NHS Trusts should also earnestly be examined (Thomas & Ashcraft, 1991).

Improved and more comprehensive quality measures would be extremely useful as physicians may very well argue that they are less efficient (take longer with patients, have longer waiting lists and so on) because they are providing better patient care. Quality variables relating to patient outcomes such as successful operations, diagnoses, morbidity and mortality rates or QALYs gained would be very useful to include in such analyses.

Finally, longitudinal data may be useful to highlight changes in efficiency and productivity of Trusts relative to peers and relative to their own performance and may help produce more robust efficiency estimates. It may be the case that between one year and the next, a Trust's activity rises, and hence it's capacity utilisation and measured efficiency also rise. A longer term look at changes in capacity utilisation and costs will assess how progress is being made towards achieving efficiency potential. Longitudinal data would help clear up several unanswered questions such as whether some outliers are merely one-off data anomalies, whether inefficient Trusts are truly that, or have made improvements on prior performance, and more importantly whether efficiency scores jump from year to year and display inconsistency. Examining efficiency over time can also better assess the level of random 'noise'. Panel data spanning a number of years with the use of stochastic frontier analysis will also allow the estimation of a fixed effect for each hospital (Skinner, 1994). The standard error of each fixed effect could then be used to make assessments of how far each Trust differs from the 'best practice' hospital. Once robust estimates of cost differences are found these differences can be analysed and interpreted. DEA Malmquist indices can also be used to examine productivity change over time (Hollingsworth, Dawson & Maniadakis, 1999).

Specification of such models will require a number of assumptions, such as how to deal with the issue of product change over time, whether the cost function is assumed constant over time, the extent to which inefficiency of each hospital is assumed constant over time, and how quality is to be dealt with.

Better data in this respect will add a great deal to understanding Trust behaviour with respect to productivity and efficiency over time and will further help to validate the results found in this study.

REFERENCES

Banker R.D, Conrad R.F & Strauss R.P. (1986) A comparative application of data envelopment analysis and translog methods: An illustrative study of hospital production, *Management Science*, 32(1): 30-44.

Chilengerian J.A. (1994) Exploring why some physicians' hospital practices are more efficient: Taking DEA inside the hospital, in Charnes, A., Cooper, W., Lewin, A.Y. & Seiford, L.M. (Eds.), *Data Envelopment Analysis: Theory, methodology and applications*, Kluwer Academic Publishers: Boston.

Charnes A, Cooper W, Lewin A.Y & Seiford L.M. (1994) *Data Envelopment Analysis: Theory, methodology and applications*, Kluwer Academic Publishers: Boston.

Coelli T. (1996a) *A guide to DEAP version 2.1: A Data Envelopment Analysis (Computer) program*, Centre for Efficiency and Productivity Analysis, CEPA Working Paper 96/08, University of New England.

Coelli T. (1996b) *A guide to FRONTIER version 4.1: A computer program for stochastic frontier production and cost function estimation*, Centre for Efficiency and Productivity Analysis, CEPA Working Paper 96/07, University of New England.

Cohen H.A. (1970) Hospital cost curves with emphasis on measuring patient care output, in Klarman H.E (Ed.), *Empirical Studies in Health Economics*, Johns Hopkins Press: Baltimore, 279-93.

Dor A. (1994) Non-minimum cost functions and the stochastic frontier: On applications to health care providers, *Journal of Health Economics*, 13: 329-34.

Feldstein M.S. (1967) *Economic analysis for health service efficiency*, North-Holland: Amsterdam.

Feldstein M.S & Schuttinga J. (1977) Hospital costs in Massachusetts: A methodological study, *Inquiry*, 14(1): 22-31.

Feldstein P.J. (1970) Comment, in Klarman, H.E. (Ed.), *Empirical Studies in Health Economics*, Johns Hopkins Press: Baltimore, 294-6.

Fernandez-Castro A & Smith P. (1994) Towards a general non-parametric model of corporate performance, *Omega International Journal of Management Science*, 22(3): 237-49.

Greene W.H. (1993) The econometric approach to efficiency analysis, in Fried, H.O., Lovell, C.A.K. & Schmidt, S.S. (Eds.), *The measurement of productive efficiency: Techniques and applications*, Oxford University Press: New York.

Greene W.H. (1995) *Limdep Version 7.0 User's Manual*, Castle Hill: Econometric Software, Inc.

Grosskopf S & Valdmanis V. (1993) Evaluating hospital performance with case-mix-adjusted outputs, *Medical Care*, 31(6): 525-32.

Hadley J & Zuckerman S. (1994) The role of efficiency measurement in hospital rate setting, *Journal of Health Economics*, 13: 335-40.

Hollingsworth B, Dawson P.J & Maniadakis N. (1999) Efficiency measurement of health care: A review of non-parametric methods and applications, *Health Care Management Science*, forthcoming.

Hollingsworth B & Parkin D. (1998) *Developing efficiency measures for use in the NHS*, A report to the NHS Executive Northern & Yorkshire R&D Directorate, February 1998, Health Economics Group, University of Newcastle.

Kooreman P. (1994) Data envelopment analysis and parametric frontier estimation: Complementary tools, *Journal of Health Economics*, 13: 345-46.

Newhouse J.P. (1994) Frontier estimation: How useful a tool for health economics?, *Journal of Health Economics*, 13: 317-22.

O'Neill L. (1998) Multifactor efficiency in data envelopment analysis with an application to urban hospitals, *Health Care Management Science*, 1: 19-28.

Orme C & Smith P.C. (1996) The potential for endogeneity bias in data envelopment analysis, *Journal of the Operational Research Society*, 47: 73-83.

Parkin D. & Hollingsworth B. (1997) Measuring production efficiency of acute hospitals in Scotland, 1991-4: Validity issues in data envelopment analysis, *Applied Economics*, 29: 1425-33.

Skinner J. (1994) What do stochastic frontier cost functions tell us about inefficiency? *Journal of Health Economics*, 13: 323-28.

Smith P.C. (1997) Model misspecification in Data Envelopment Analysis, *Annals of Operations Research*, 73: 233-52.

Smith P.C. (1998) *Data envelopment analysis in health care: An introductory note*, Centre for Health Economics: University of York.

Söderlund N. & van der Merwe R. (1999) *Hospital benchmarking analysis and the derivation of cost indices*, Centre for Health Economics Discussion Paper No. 174, University of York.

Thanassoulis E. *Warwick DEA User Manual*, Warwick Business School: Warwick University.

Thomas J.W & Ashcraft M.L.F. (1991) Measuring severity of illness: Six severity systems and their ability to explain cost variations, *Inquiry*, 28: 39-55.

Valdmanis V. (1992) Sensitivity analysis for DEA models: An empirical example using public vs. NFP hospitals, *Journal of Public Economics*, 48: 185-205.

Vitaliano D.F & Toren, M. (1994) Frontier analysis: A reply to Skinner, Dor and Newhouse, *Journal of Health Economics*, 13: 341-43.

Wagstaff A. (1989) Estimating efficiency in the hospital sector: A comparison of three statistical cost frontier models, *Applied Economics*, 21: 659-72.